Experimental Analysis of a Spatialised Audio Interface for People with Visual Impairments

JACOBUS C. LOCK*, University of Lincoln, UK
IAIN D. GILCHRIST, University of Bristol, UK
GRZEGORZ CIELNIAK, University of Lincoln, UK
NICOLA BELLOTTO, University of Lincoln, UK

Sound perception is a fundamental skill for many people with severe sight impairments. The research presented in this paper is part of an ongoing project with the aim to create a mobile guidance aid to help people with vision impairments find objects within an unknown indoor environment. This system requires an effective non-visual interface and uses bone-conduction headphones to transmit audio instructions to the user. It has been implemented and tested with spatialised audio cues, which convey the direction of a predefined target in 3D space. We present an in-depth evaluation of the audio interface with several experiments that involve a large number of participants, both blindfolded and with actual visual impairments, and analyse the pros and cons of our design choices. In addition to producing results comparable to the state-of-the-art, we found that Fitts's Law (a predictive model for human movement) provides a suitable a metric that can be used to improve and refine the quality of the audio interface in future mobile navigation aids.

CCS Concepts: • **Human-centered computing** → HCI theory, concepts and models; *Laboratory experiments*; **Sound-based input / output**; Pointing devices; **Empirical studies in accessibility**.

Additional Key Words and Phrases: Visual impairment; active vision; guidance system; audio interface; Fitts Law

ACM Reference Format:
Jacobus C. Lock, Iain D. Gilchrist, Grzegorz Cielniak, and Nicola Bellotto. 2020. Experimental Analysis of a Spatialised Audio Interface for People with Visual Impairments. 1, 1 (July 2020), 22 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The ActiVis project is an effort to create a fully mobile phone-based navigation aid for people with visual impairments to guide them towards a desired object or location within an unknown indoor environment. Improvements to computer vision and machine learning techniques, as well as mobile computing hardware performance, are exploited to make this system possible. In particular, techniques from the active vision field are used to enable a mobile device to gather information on the surrounding environment and use it to generate guidance instructions for a user with limited

*Contact Author

Authors' addresses: Jacobus C. Lock, jaycee.lock@gmail.com, University of Lincoln, Brayford Pool, Lincoln, UK, LN6 7TS; Iain D. Gilchrist, University of Bristol, Bristol, UK; Grzegorz Cielniak, University of Lincoln, Lincoln, UK; Nicola Bellotto, University of Lincoln, Lincoln, UK.

XXXX-XXXX/2020/7-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

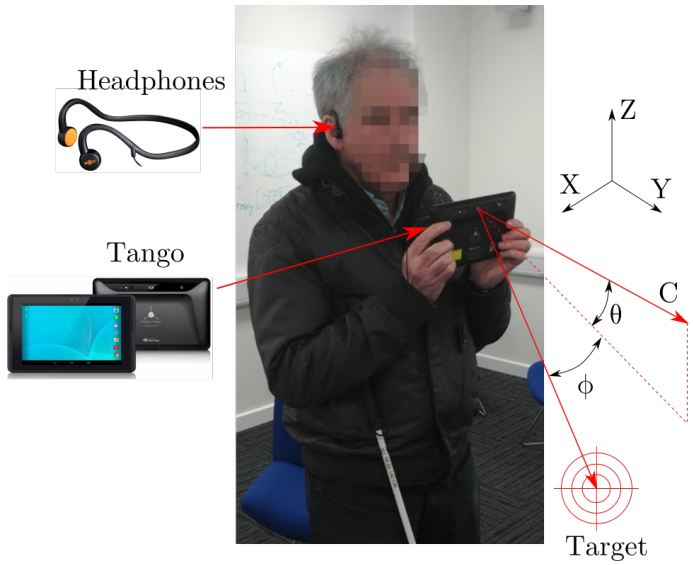

Fig. 1. The Tango tablet and bone-conduction headphones in use during an experiment. The angle reference system that the guidance instructions are based on is also included , showing the camera vector, C, and the pan and elevation angles to the target. Adapted from Lock et al. [2019b].

or no vision. However, in the literature, these techniques are typically limited to directing electro-mechanical servos [Bajcsy et al. 2018] and in this project, we attempt to find out if similar techniques can be used to direct a user's attention to a target location. The proposed system was introduced by Lock et al. [2017], who describe the aforementioned autonomous guidance system paired with a co-adaptive human-machine interface that changes its own parameters over time to better match the user's perception strengths and limitations. In previous work we developed a prototype guidance system that uses active vision and machine learning models to gather information and help a person find objects within an unknown indoor environment, showing that these techniques can indeed successfully be applied to direct humans' attention [Lock et al. 2019a]. However, in this work we examine the effectiveness of the proposed interface for a searching task and investigate a metric that can be used in the next phases of the project to enable the aforementioned co-adaptive paradigm that could benefit the user experience and boost long-term navigation performance.

The hardware used for the guidance system prototype are a Google Project Tango device[1], along with a set of bone-conduction headphones (both are shown in Fig. 1). This Tango device uses a set of embedded RGB-D cameras and other hardware and software components to provide accurate real-time localisation estimates and has powerful image-processing capabilities. It also gives access to Android's full range of interface and I/O options. Unlike normal headphones, bone-conduction headphones are placed on and conduct audio signals through a user's cheekbones, instead of their ear canals. These headphones have the benefit of not interfering with a user's normal hearing function.

The system generates guidance instructions that can be interpreted in real-time with minimal additional cognitive load to the user, given humans' natural ability to determine a sound source's 3D position. By adjusting a tone's spectral make-up (elevation angle), time delay and level difference (pan angle), and intensity (distance), a sound source can be spatialised to come from any arbitrary

[1]https://en.wikipedia.org/wiki/Tango_(platform)

location. In this case, only the pan and elevation angles are given to the user to instruct them to point the camera towards a target object or visual feature. However, the spectral signature generated by bone-conduction headphones cannot properly be interpreted by a human, since they are placed on the user's cheekbones and bypass the ear's outer structure. We address this limitation by having the system convey the target's elevation angle by adjusting the tone's pitch instead of spatialising it in the elevation dimension. In previous work, we showed that this audio signal transmission scheme can direct users to a target position with a level of accuracy comparable to fully spatialised signals used with expensive closed-cup headphones. These results were presented at the 9th International Workshop on Assistive Engineering in 2019 [Lock et al. 2019b]. We expand upon this initial investigation with a larger dataset and look at if, and how, changing the behaviour of the pitch affects target acquisition performance in terms of time and angular error. The participants' hearing characteristics are also measured to determine if there are limitations to how well they can determine audio pitch or direction.

The main contributions of this paper are two-fold:

- we provide comprehensive experimental results, with two groups of participants with healthy and limited eyesight, on how well a tone with varying pitch can convey a target's elevation angle when using a mobile device with a bone-conduction headset;
- we show that this sound-based human-machine interface exhibits a response that is well-modelled by Fitts's Law which can provide a useful metric of performance for similar mobile user interfaces.

The rest of the paper starts by discussing relevant works and previous research in Section 2. This is followed by a discussion on the design and implementation of our interface in Section 3. The experiments that were conducted are discussed in Section 4, while their results are presented and discussed in Section 5. The paper concludes with a summary of the work and discussions on future research prospects in Section 6.

2 PREVIOUS WORK

Over the years, commercial and academic groups have devised new and innovative mobile navigation and electronic travel aids (ETA) for people with visual impairments to address macro- and micro-navigation issues. The latter refers to the scale in which navigation takes place, with macro-navigation addressing directional guidance on a topographical map using turn-by-turn instructions, for example, while micro-navigation focusses on conveying information of the user's immediate surroundings [Petrie et al. 1997]. Many of the proposed systems make use of one or a combination of vocal [Chessa et al. 2016; Kanwal et al. 2015; Mocanu et al. 2016; Sato et al. 2019], audio [Kammoun et al. 2012; Rodríguez et al. 2012; Schwarze et al. 2015] and haptic [Lee and Medioni 2015; Rivera-Rubio et al. 2015; Xiao et al. 2015] feedback media to present a user with macro- and/or micro-guidance instructions, each of which has its own set of features and limitations. The interface for this work is used to provide instructions to a stationary user to guide them to point a camera at a target location and we therefore are mainly concerned with micro-navigation tools. In this context, respondents with visual impairments typically prefer haptic and vocal feedback over an audio tone [Arditi and Tian 2013; Golledge et al. 2004]. However, from an ergonomic perspective, haptic feedback has significant shortcomings and require additional external hardware to transmit guidance instructions with sufficient resolution. Furthermore, in a micro-navigation task, high-resolution guidance and many adjustments are required to reach the target location. In this regard, both haptic and vocal feedback can present a cognitive burden and overwhelm a user's input bandwidth, which could have a detrimental effect on performance and the overall user experience [Klatzky et al. 2006]. As an alternative, simple audio tones are less prone to these

bandwidth and hardware limitations. However, such tones can potentially fatigue a user if they are too unpleasant. Researchers have started using cameras and object detectors to determine what a user is looking at and then use simple audio tones and vocal feedback signals to indicate to the user where to find an object. They report favourable results with these systems [Fiannaca et al. 2014; Schauerte et al. 2012; Tian et al. 2013; Vázquez and Steinfeld 2012].

Work has been done in an attempt to spatialise these tones with a head-related transfer function (HRTF) that simulates a sound source being placed at some arbitrary 3D position. The authors generally report favourable results when using normal over-ear headphones or speakers [Blum et al. 2013; Crispien and Petrie 1994; Geronazzo et al. 2016; Katz et al. 2010; Presti et al. 2019; Wilson et al. 2007]. However, the audio transmission device used has a significant effect on performance. Indeed, research has shown that cheaper headphones and bone-conduction headphones have diminished performance in location perception with HRTFs when compared to over-ear or other expensive headphones [Mascetti et al. 2016; Stanley and Walker 2006; Voong and Oehler 2019]. However, this performance degradation seems limited to the elevation dimension, which is drastically improved when the HRTFs are adjusted for the bone-conduction pathway [Pec et al. 2008; Stanley and Walker 2006]. Another way around the performance issue in the elevation dimension is to transmit the elevation angle by adjusting the audio tone's pitch, such as in the work by Durette et al. [2008]. We expanded upon the latter's work and investigated using spatialised audio and a varying pitch for guidance in an initial study [Lock et al. 2019b], and found that participants are able to adequately determine a target's elevation. The measured performance is comparable to those of more expensive and over-ear headphones.

Fitts's Law [Fitts 1954] is a predictive model of human movement and is particularly useful to evaluate human-computer interactions. Indeed, researchers have previously used Fitts's Law, and more recently MacKenzie's modified version of it [MacKenzie 1992], as a metric to evaluate the performance of a spatial audio HMI. Fitts's Law was originally proposed for visual target search tasks, but has since been applied in non-visual target search tasks as well. For example, experiments with a haptic feedback pointing device have been performed to evaluate how effective it was at directing a user towards a target [Ahmaniemi and Lantz 2009] and the authors showed that the search time adheres to Fitts's Law. However, they also note that it is not a perfect fit, citing the fact that Fitts's Law does not take into account a user's search strategy. Another group of researchers conducted experiments using a spatial audio interface to describe the position of a target on the horizontal plane [Marentakis and Brewster 2006]. Here, participants pointed to where they thought the targets were, on their left or right, as they traversed a path. Their results show a good relation between target difficulty and search time, providing a strong argument that Fitts's Law can be used to describe the performance of a spatial audio interface. These results have since been supported by other authors, who found that Fitts's Law provided a good explanation for the results from an experiment using visual, limited visual and non-visual feedback cues [Wu et al. 2010]. However, it is not clear whether Fitts's Law applies to a spatial tone that uses varying pitch to convey the target's elevation angle, as demonstrated in this paper.

3 SYSTEM DESCRIPTION

Existing electronic navigation aids have typically struggled to gain significant market traction and replace the traditional walking cane as the standard assistive tool for people with visual impairments. Current technological limitations include prohibitive costs, bulky hardware requirements and non-user-friendly interfaces [Arditi and Tian 2013; Golledge et al. 2004; Yusif et al. 2016]. To address these issues, we implemented a handheld mobile system that is based on a concept proposed by Lock et al. [2019a] and tested by Lock et al. [2019b] using a Google Tango device that is able to localise

itself in real-time. This system has the benefit of minimal hardware requirements and a compact, familiar form-factor, which will help to overcome the hurdle of user-acceptance and usability.

People with visual impairments rely heavily on their hearing [Golledge et al. 2004] and we wish to avoid blocking this information pathway, since doing so can have severe and undesirable effects [Lichenstein et al. 2012]. We therefore use a set of bone-conduction headphones that are placed on a user's cheekbones and conduct audio signals through the skull to the inner ear, completely avoiding the outer ear and does therefore does not impede the user's ambient sound perception abilities. Alternative, open-back headphones were also considered, but it was found that they still interfere with hearing and were therefore disregarded. The AfterShockz headphones (shown in Fig. 1) were ultimately selected. These headphones have the added benefit of a more discreet form-factor when compared to other over-headphones, thereby addressing the issue of user acceptance.

Humans are able to localise sounds in three dimensions by extracting different cues from audio signals [Blauert 1969, 1997]. These include binaural, where the user compares the signals received at both ears (e.g. inter-aural time and level differences), and monaural cues, where a cue is extracted from the signal received at each ear (e.g. the spectral profile, audio intensity). Prior to transmitting an audio signal to the user, it can be adjusted by an HRTF to mimic a natural sound source and make a user believe it is located at some arbitrary 3D location. An HRTF is a mathematical function that simulates a human head and ear's response to an external sound and is derived by capturing key characteristics that affect the monaural and binaural responses at the user's ear, such as the hearing level and anatomy. An HRTF will produce the most accurate results when it is customised to match a specific user, since each user's hearing response is unique, but this is a complex process and HRTFs generated with average values (head sizes, heights, ear shapes, etc.) produce acceptable results [Gardner and Martin 1995].

The interface presents the user with guidance instructions in the form of angular adjustments that are required in the pan and elevation dimensions to point a camera at the desired target (see Fig. 1). Spatialised audio is well-suited to this micro-navigation task, displaying accuracy comparable to vocal feedback, but with less cognitive load when used in a high-resolution search task such as this [Klatzky et al. 2006]. However, we propose a linear adjustment to the signal's pitch as a function of the elevation angle to overcome the limitations of bone-conduction headphones and spatialised audio stated previously. The pan angle can still be conveyed with a spatialised audio signal generated by an HRTF. Indeed, this dimension is unaffected by the use of bone-conduction headphones [Lock et al. 2019b; MacDonald et al. 2006; Schonstein et al. 2008; Stanley and Walker 2006]. This interface was implemented and evaluated in Lock et al. [2019b].

3.1 Pan Dimension

To localise a sound on the horizontal plane, the human audition system compares characteristics from the signals received at both ears (binaural cues), such as their volume difference and the time delay between the same sound reaching both ears [Blauert 1969]. For example, a sound placed at a person's right will hit their right ear first with a slightly higher volume compared to the left ear. Since the binaural cues are largely independent if the signal's spectral profile, it would be convenient to use a simple audio wave to transmit guidance instructions to the user. Therefore, in this work a pure sine wave was used, but can easily be replaced with a richer tone once the system's characteristics are better understood. The since was was spatialised using OpenAL's default HRTF, based on the MIT's KEMAR dataset [Hiebert 2005].

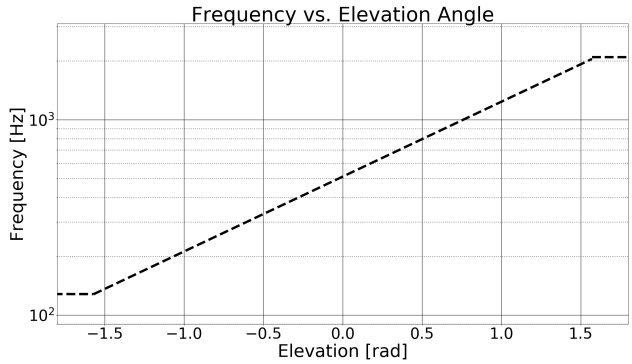

Fig. 2. A plot depicting the pitch gain function used to convey the target's elevation angle. Note the logarithmic scale of the frequency axis. Adapted from Lock et al. [2019b].

3.2 Elevation Dimension

To compensate for the loss of elevation localisation performance when conveying spatialised audio via bone-conduction headphones [MacDonald et al. 2006; Schonstein et al. 2008], we adjust the sine wave's frequency (i.e. audible pitch) as a log-linear function of the target's elevation angle, as shown in Fig. 2. When the required elevation adjustment is above or below where the camera is currently pointed, the pitch is increased or lowered respectively. This high/low scheme was selected based on humans' natural association of high-pitched sounds with elevated objects and vice-versa for low-lying objects [Blauert 1997; Pratt 1930]. The pitch is constantly adjusted and updated at 10 Hz in octave- and semitone-based intervals to ensure perceptible changes, while keeping the tone's timbre constant [Shepard 1964].

The pitch is changed as a linear function of the elevation angle, the gradient of which is determined by setting the angle and pitch limits. These limits are set at some number of octaves from the neutral pitch that is emitted when the camera is on-target. This tone is heuristically set to 512 Hz following practical tests, which is comfortably audible for a large number of octaves changes. For this work, we only consider a 180° field of view in front of the user and limit the elevation angle to a range of ±90°, or $[-\frac{\pi}{2}, \frac{\pi}{2}]$. In previous tests with this specific interface, its performance was comparable to normal over-ear headphones transmitting a signal fully spatialised in the pan and elevation dimensions [Lock et al. 2019b].

4 EXPERIMENTS

We performed a set of experiments with the audio interface to determine how effective it is at directing a user to adjust the pan and elevation angles of a camera to point it to a target. Furthermore, we also carried out a set of pre-screening experiments to determine each participant's hearing characteristics in order to determine their perception limits in the respective audio dimensions. The participants were given time before the experiments commenced to familiarise themselves with the device and the tones it emits, as well as what the 'on-target' tone sounds like. We also tested the system with a group of participants with severe sight impairments and compared their data to the blindfolded participant dataset. The results from the experiments we performed allow us to better understand how the users respond to different settings for the spatial audio feedback stimulus and use them to improve and optimise the behaviour of the feedback modes in our portable navigation aid.

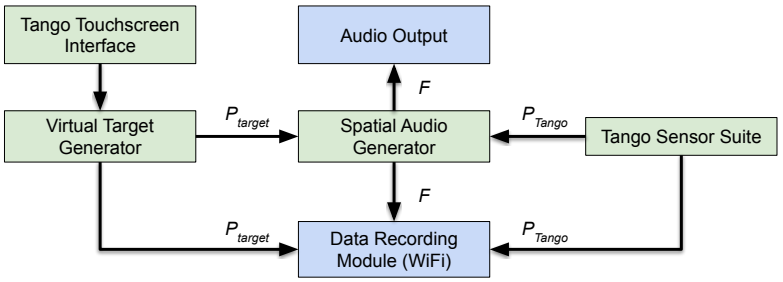

Fig. 3. A diagram of the individual system components and their communication pipelines. F indicates a feedback signal and P a pose signal. Adapted from Lock et al. [2019b].

4.1 System Setup

A diagram of the experimental system pipeline is shown in Fig. 3, where the arrows indicate the direction of information flow. The system implementation used here is similar to the setup used by Lock et al. [2019b]. When the user taps the Tango's screen, a new virtual target is generated and its coordinates are sent to the audio generation module, along with the device's current position and orientation. The audio generator then produces a tone based on the difference between the device and the target's positions. The tone is sent to the audio output channel, which plays it back to the user. A WiFi recording module is constantly monitoring the different values of the device's parameters and of the target's position, as well as the system's output, recording everything in a remotely stored datafile.

Even though the Tango is able to detect distances to object, we opted to use virtual targets given the added ability to place targets at any random location without physical manipulation without losing the ability to determine its exact position relative to the device. When this interface is used with a real navigation system, it will be extended to work with real targets and objects.

4.2 Participant Characterisation

A preliminary set of experiments were conducted to characterise the participants' hearing characteristics. The measured characteristics were each participant's audio localisation ability on the lateral plane, as well as the participants' ability to discriminate between tones with different frequencies. These results will be used to provide context to the following target search experiment and provide additional insight on any possible biases or limitations.

4.2.1 Sound Localisation. In this experiment, we evaluated a participant's ability to determine the lateral direction a sound is coming from. To do this, we played a continuous 512 Hz sinusoidal tone to the participant through the headphones and applied an HRTF to spatialise and place its source to the participant's left or right. The participant then had to select the direction the sound came from. The longer the experiment lasts and the more correct guesses the participant makes, the closer the source moves to the centre-front of the participant, making it increasingly harder to localise.

For this progressive increase in difficulty, a "2-up, 1-down" step process is used [Levitt 1971; Wetherill and Levitt 1965], meaning that for every two correct answers, the distance to the centre halves. Conversely, the task becomes easier for each incorrect answer by doubling the sound source's distance from the centre. We also use two different step sequences, one starting at a large angular distance (45°) from the user and the other at the minimum distance (approximately 1°), giving an 'easy' and a 'hard' progression respectively. The terminating condition for the experiment

is when the two sequences converge to within two intervals of one another for three consecutive guesses. For example, the experiment will terminate when one sequence is set to 11.25° and the other is between 2.8° or 45° for three consecutive guesses. This gives a distance band where the participant is capable of localising the sound source. Each participant performed this experiment three times.

4.2.2 Pitch Discrimination. Here we determined a participant's ability to differentiate the pitches of two different tones, i.e. how well they can tell if a tone is high or low pitched. We played two tones to the participants in succession, with the second tone being higher or lower-pitched than the first. The participants were then asked to select which tone was higher or lower.

The first tone is randomly generated, while the second tone is generated by adding or subtracting some value from the first one. The tone difference depends on how well the participant can tell the tones apart. Like the sound localisation experiment, a "2-up, 1-down" step process is used: for every two consecutive correct answers, the pitch difference between the tones is halved, while it is doubled for every incorrect answer. Two-step sequences are again used here, one starting with a large pitch difference ($f_h = 2^9 = 512\,\text{Hz}$) between the tones and the other with a small difference ($f_l = 2^1 = 2\,\text{Hz}$). The termination condition is when the two-step sequences are within one octave of each other (i.e. $\log_2 \frac{f_h}{f_l} = 2$) for three consecutive answers. For example, the experiment will terminate when one sequence is set to 64 Hz and the other is between 32 Hz or 128 Hz for three consecutive guesses. Pitch differences are measured in semitones, which can be obtained with

$$\Delta f = 12 \log_2 \frac{f_0}{f_1}, \tag{1}$$

where f_0 and f_1 are the frequencies of the first and second tone respectively. Each participant performed this experiment twice.

4.3 Target Search

A set of experiments were conducted to determine the interface's guidance effectiveness for a pointing task. These experiments captured the difference between the targets' actual directions and the directions the participants perceived them to be located. The participants were given a Tango tablet running an app implementing the experimental setup in Fig. 3. This app generates a set of virtual targets, one at a time, and presents their directions to the participants with audio guidance signals. These targets are generated at a constant distance from the participants with pan and elevation angles that are uniformly distributed across the four quadrants to avoid clustering. Each target's relative angular position is communicated to the participants in real-time as the device is moved around via the bone-conduction headphones. Every time a participant was confident that they were pointing at a target, i.e. hearing a signal that is placed at their front at a frequency of 512 Hz, they tapped the device's screen, marking the location and generating a new target. The targets are all positioned relative to the device's camera coordinate system, which is tracked by the Tango hardware and localisation API. Each participant was tasked with finding 28 targets each.

A similar experimental setup was used in [Lock et al. 2019b]. However, in these experiments we also want to see how changing the gradient of the pitch function, visualised in Fig. 2, affects target acquisition performance, e.g. does a steeper pitch gain as a function of the elevation angle improve accuracy or decrease the search time? Pitch limits of one, two and three octaves above and below the neutral tone were then set for the so-called *lo*, *med* and *hi* pitch gradient settings respectively, giving pitch intervals of

$$f_{lo} \in [256\,\text{Hz}, 1024\,\text{Hz}]$$
$$f_{med} \in [128\,\text{Hz}, 2048\,\text{Hz}]$$
$$f_{hi} \in [64\,\text{Hz}, 4096\,\text{Hz}].$$

We use two different metrics to compare the three different pitch gradient settings: acquisition accuracy and search time. The accuracy is given as the difference between the Tango's orientation at the time the participant confirmed they were on target, and the target's actual orientation. We separate the results of the elevation and pan dimensions in order to see how the different pitch gradients affect a participant's pointing accuracy.

We also compare the performance of the three pitch gradient settings in terms of the time it takes each participant to find a target. However, since each participant was presented with a different, randomly generated set of targets, a direct time comparison is not possible. Instead, we use Fitts's Law [Fitts 1954], modified by MacKenzie for uncertain target sizes and noisy data [MacKenzie 1992], which states that there is a relation between the time it takes to find a target and its index of difficulty (*ID*, the ratio between the distance to the target and its width). It also provides a so-called 'index of performance' (*IP*), which can be used to compare the results between the three different configurations used in the experiments. Furthermore, if a Fitts-like relationship is indeed found, it can be used as an optimisation metric in the co-adaptive interface described by Lock et al. [2017], where interface parameters (e.g. pitch gradient, volume setting, etc.) are automatically tweaked in order to maximise target finding performance.

Here we briefly summarise the equations and quantities involved in this metric. Fitts's Law is given by

$$t = a + b\,ID, \tag{2}$$

where t is the time it takes to find a target, a and b are constants determined through regression and ID is a description of the difficulty of the target, given as logarithmic function of the ratio between the distance to the target and the target's width. In our case, the targets have no width, since they are points in space, and we therefore use MacKenzie's modified form for ID, given by

$$ID = \log_2\left(\frac{\theta}{w_e} + 1\right). \tag{3}$$

Here θ is the angular distance between subsequent target centres and w_e is the targets' effective angular width [Welford 1968], given by

$$w_e = \sqrt{2\pi e}\,\sigma = 4.133\,\sigma, \tag{4}$$

where σ is the standard deviation of the combined pan and elevation error (x, y), calculated with

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}\left(\sqrt{(x_i - \bar{x})^2 + (y_i - \bar{y})^2}\right)^2}{N-1}} \tag{5}$$

as proposed by Wobbrock et al. [2011], taken as the angle between the participant's target selection and target's actual angular position. The virtual targets have a programmed radius of approximately 10 cm (equivalent to approximately 0.1 rad) and when the participant pointed the camera within this radius, 512 Hz on-target tone was emitted. This is similar to the work by Kabbash and Buxton [1995], where they had participants search for a point-target with an oversized cursor and elicited a Fitts-like response. However, in this case the participants were not explicitly informed when they

Table 1

	Group *G1*	Group *G2*
Gender [M/F]	10/32	7/3
Age [years]	20 ± 2	61 ± 17
Degree of Vision Impairment	N/A	7 totally blind, 3 with very limited light perception
Experience with ETAs	None	None

were within this radius and had to use their own subjective judgement of the audio tone. Since the judgement accuracy will vary across different participants, effectively giving different target sizes, it was necessary to use the effective width as an approximation for the actual target size. Previous authors have found this to capture the true performance metrics fairly well [Zhai et al. 2004]. Fitts's index of performance, IP, can then be calculated using

$$IP = \frac{ID}{t}. \tag{6}$$

4.4 Procedure

Two groups of participants were recruited for the experiments on a volunteer basis. Group *G1* consisted of 42 undergraduate students (10 male, 32, female) with normal eyesight who were blindfolded for the experiments (mean age: 20 ± 2 years). Group *G2* contained 10 people (7 male, 3 female) with severe visual impairments (mean age: 61 ± 17 years). Of the latter group, 3 are congenitally blind, while the rest were classified as severely sight impaired later in life. Of these, 3 participants still have limited light perception with no ability to reliably discern shapes and objects (the rest had no light perception). Nevertheless, they were asked to close their eyes during the experiment. None of the participants reported any significant prior experience with electronic navigation aids and none had any hearing or other disabilities that could have influenced their performance in the experiments. These demographics are summarised in Table 1.

Each participant performed three sets of experiments each, with the two characterisation experiments in Section 4.2 preceding the final target-search experiment in Section 4.3. Both groups were given some time before the target-search experiment to familiarise themselves with the system, the audio signal's behaviour and the 512 Hz on-level tone. Furthermore, to minimise any potential speed/accuracy biases, we asked the participants to focus on finding the targets without worrying about the time it took to complete the task.

5 RESULTS

5.1 Characterisation of Sound Localisation

Fig. 4 shows the results captured from the sound localisation experiment where the participants had to select the direction (left or right) that the tone was being played from. It can be seen that the vast majority of guesses for both groups were correct. For Group *G1*, most of the errors were made at the minimum distance from the centre, i.e. the most difficult to guess correctly, which is the expected behaviour. This indicates that the participants in *G1* consistently progressed through the distance intervals and we can therefore conclude they had little difficulty determining sound direction.

Group *G2* also displays a concentration of erroneous guesses in the central interval. However, it also shows more errors in other distance intervals and a more even progression towards the centre.

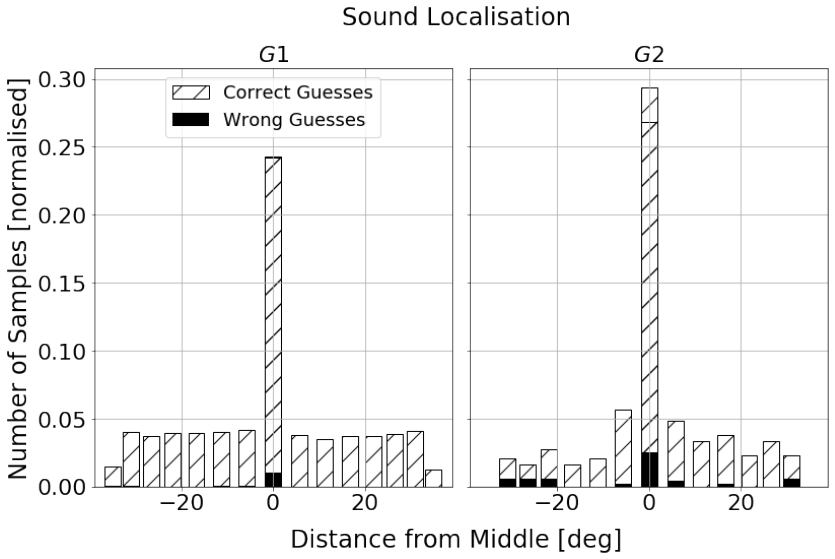

Fig. 4. Histograms of the participants' guesses of the tone locations that show the correct and incorrect guesses for each bin.

This could indicate that, instead of terminating the experiment as described in Section 4.2.1, there was more switching back and forth between the three central intervals. These results show that both participant groups are capable of determining a sound source's location with a reasonable level of consistency and accuracy and are in line previous literature, confirming that humans are very adept at localising a sound source, particularly in the pan dimension [Wersenyi 2003].

5.2 Characterisation of Pitch Discrimination

The results of the pitch discrimination experiment are shown in Fig. 5, where bar plots are used to show the proportion of correct to incorrect guesses of which tone was higher pitched for different tone difference intervals. For Group *G1*, we see that their guesses are normally spread around the 0 semitone-difference interval and the highest proportion of incorrect guesses occurs in the $[-0.25, 0.25]$ semitone-difference interval. The guesses from Group *G2* are more concentrated around the centre and the majority of incorrect guesses also occurs in the $[-0.25, 0.25]$ semitone-difference interval. The concentration of incorrect guesses around the centre is expected, given the experiment process's progressive increase in difficulty.

Assuming these differences are normally distributed, we fit a cumulative distribution function (CDF) over each participant's set of results for their correct guesses. We then used each CDF's parameters to determine a frequency cut-off threshold, where the participant could not longer reliably tell tones apart, which is set to contain 75% of each participant's correct guesses starting from the easiest to distinguish tones with large frequency differences, to the hardest to distinguish with the smallest differences. The median of these threshold values can then be used to estimate the frequency difference at which the entire participant population can no longer tell the difference between two tones. It can also be used to improve the interface's frequency profile and performance. Fig. 6 shows the threshold distribution, along with the median value, which was found to be 0.29 for Group *G1* and 0.35 for Group *G2*.

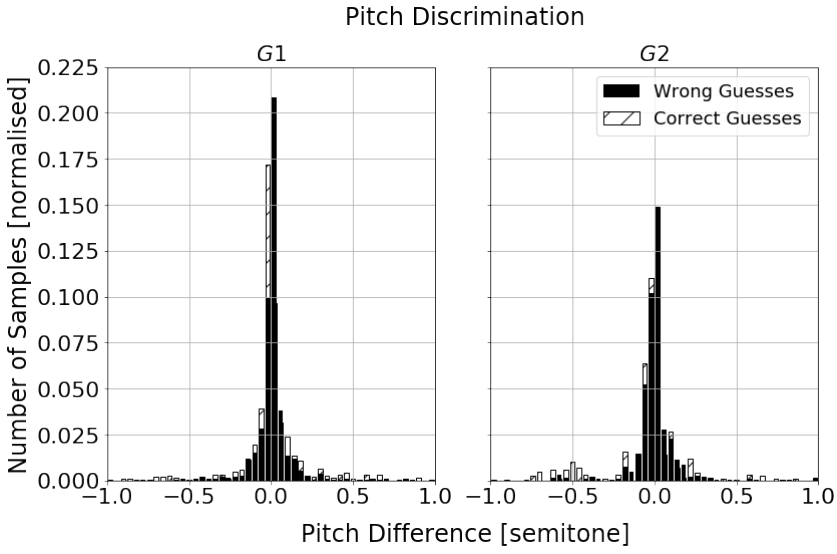

Fig. 5. Histograms of the participants' guesses of which tone was higher pitched that show the correct and incorrect guesses for each bin.

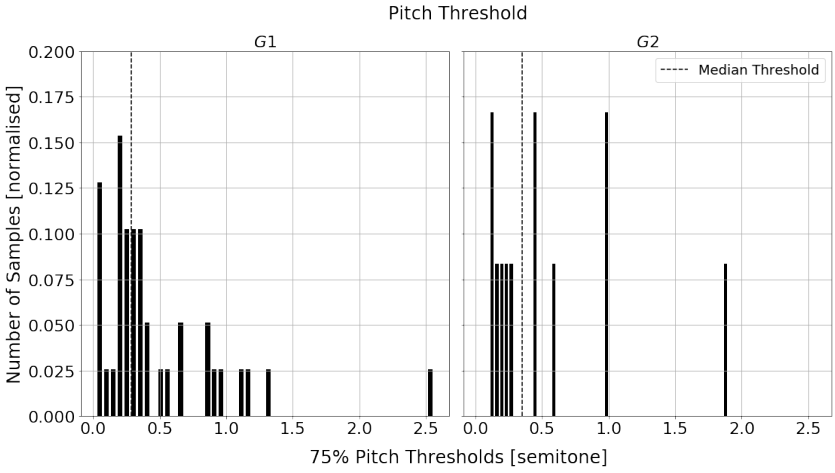

Fig. 6. Distributions of the median cut-off frequency thresholds along with the median 75% cut-off thresholds.

For the target search experiment, the pitch differences between the 512 Hz on-target tone and the participants' selected tones were collected and their cut-off threshold were determined in a similar way. Each participant's median tone error for each setting was then plotted alongside the groups' median thresholds in Fig. 7. These data are plotted on a linear Hertz-based scale instead of a semitone scale to highlight the grouping between the three settings.

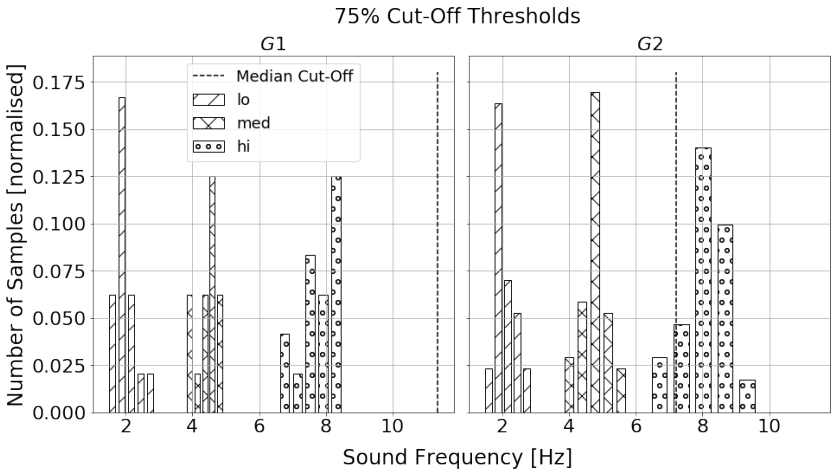

Fig. 7. Histogram distributions of the participants' 75% cut-off thresholds.

Table 2. The average target acquisition error in the pan and elevation dimensions for each participant group.

		Setting	Mean Angle Error [rad]	Mean Absolute Angle Error [rad]	Pearson Correlation
G1	Pan	*lo*	-0.02 ± 0.37	0.25 ± 0.27	0.75, $p < 0.001$
		med	-0.01 ± 0.37	0.26 ± 0.27	0.77, $p < 0.001$
		hi	-0.03 ± 0.39	0.26 ± 0.29	0.72, $p < 0.001$
	Elevation	*lo*	-0.12 ± 0.51	0.42 ± 0.31	0.36, $p < 0.001$
		med	-0.11 ± 0.41	0.44 ± 0.24	0.49, $p < 0.001$
		hi	-0.15 ± 0.44	0.36 ± 0.29	0.48, $p < 0.001$
G2	Pan	*lo*	-0.01 ± 0.37	0.48 ± 0.31	0.10, $p = 0.03$
		med	0.04 ± 0.53	0.45 ± 0.27	0.13, $p = 0.01$
		hi	0.03 ± 0.48	0.36 ± 0.22	0.21, $p < 0.001$
	Elevation	*lo*	-0.30 ± 0.59	0.49 ± 0.39	0.03, $p = 0.48$
		med	-0.42 ± 0.45	0.42 ± 0.33	0.31, $p < 0.001$
		hi	-0.37 ± 0.43	0.36 ± 0.32	0.40, $p < 0.001$

5.3 Target Search

The results from the target search experiment shown on the 2D histograms in Fig. 8, where the angular errors in the pan and elevation dimensions are plotted against each other. A set of box-plots of the angle errors are also given in Fig. 9 for each audio setting. The results are summarised in Table 2.

The Shapiro-Wilkes test for normality reveals that none of these distributions are normally spread. Therefore, the Pearson test is used to investigate the correlation between the actual target location and participants' pointing location. These results are included in Table 2. The Pearson correlation scores for Group *G1* indicate a moderate to strong positive correlation between the target and the selected locations ($r_{pan} \in [0.72, 0.77],\ p < 0.001;\ r_{elevation} \in [0.36, 0.49],\ p < 0.001$), showing that both the pan and elevation cues in general worked as expected. However, the correlation scores for

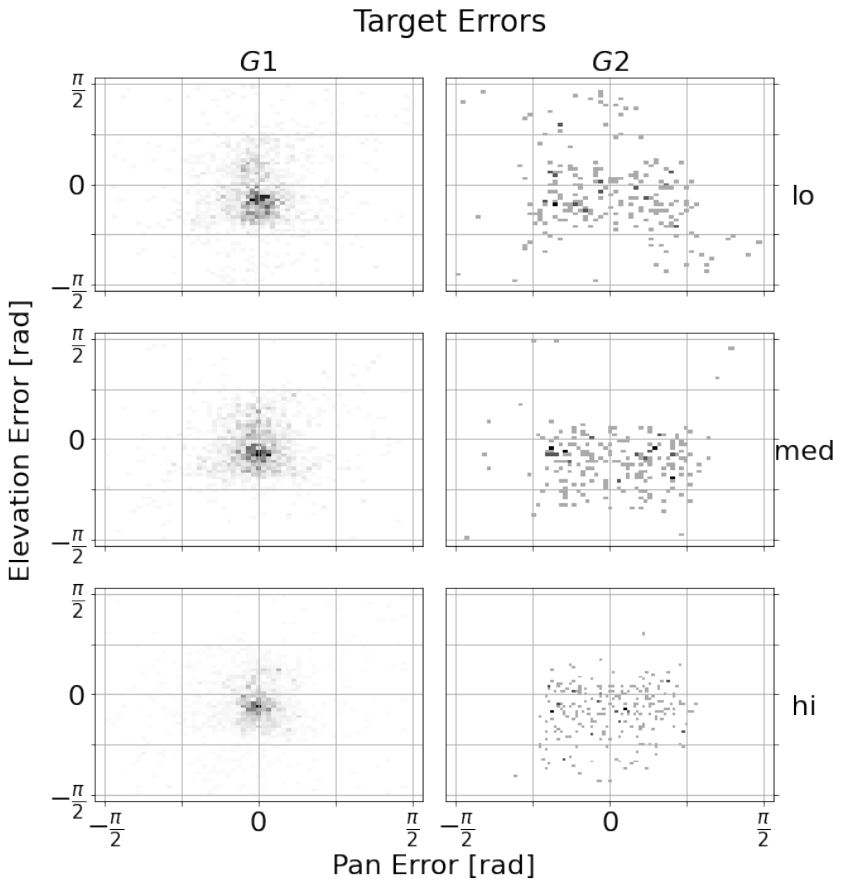

Fig. 8. Distributions of the angular errors in the pan and elevation dimensions for the 3 different pitch gradient settings.

Group *G2* are significantly weaker, with a pan angle correlation $r_{pan} \in [0.1, 0.21]$, $p < 0.03$. With the exception of the *lo* setting ($p_{lo} = 0.48$), the elevation correlation is generally stronger, with $r_{elevation} \in [0.31, 0.40]$, $p < 0.001$.

The repeated-measures procedure that was used for these experiments requires the data for each participant to be grouped together for each setting. The medians of these data groupings are then used as individual samples that represent an individual participant's performance for each setting. Fig. 9 shows these median data collected from each participant as a set of box-plots, while Fig. 10 shows the collection of absolute errors.

The box-plots in Fig. 9 show that the error in the pan dimension is approximately centred around 0 rad for both groups, with some divergence between the groups for the different settings. However, using the Friedman test for repeated measures on the medians of absolute errors, these divergences are found to be not significant ($p_{G1} = 0.17$, $p_{G2} = 0.09$), showing that spatial perception and accuracy are not affected by changes in the tone's pitch. This is further demonstrated in the box-plots in Fig. 10, which demonstrates relatively consistent error levels in the pan dimension for both groups and across all three settings.

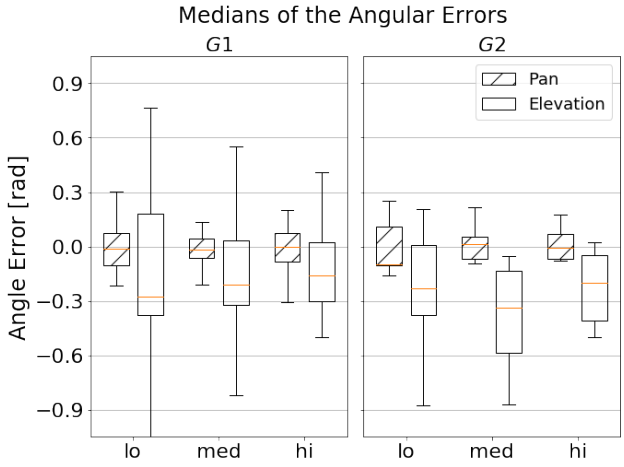

Fig. 9. Box-plots of the median pan and elevation errors for each audio setting.

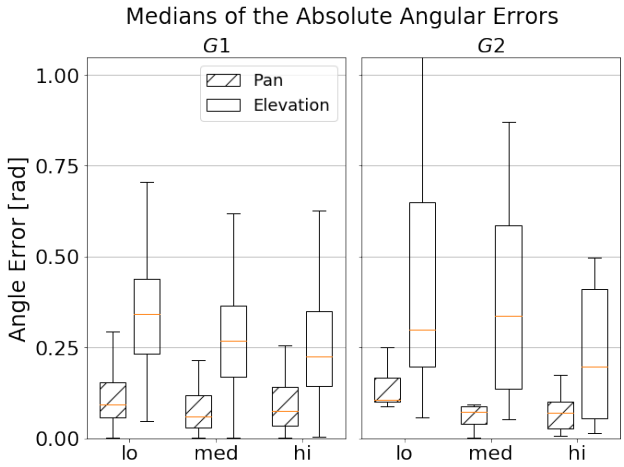

Fig. 10. Distributions of the absolute angular errors in the pan and elevation dimensions for the 3 different pitch gradient settings.

Regarding the errors in the elevation dimension in Fig. 9, we observe in Group *G1* a narrowing distribution between the *lo*, *med* and *hi* settings, respectively, and a median gradually approaching 0 rad. A similar trend is observed for Group *G2*, but the improvement across the settings is more subtle and not as linear as for Group *G1*. Fig. 10 shows a clearer improvement, i.e. approaching 0 rad, for the elevation data between the three settings in both groups, with the *hi* setting producing the smallest error in both cases. Further analysis of the medians of the absolute elevation error with the Friedman test reveals that the results for the different settings are significantly different from each other only for Group *G1* ($p_{G1} = 0.002$, $p_{G2} = 0.32$).

A post-hoc analysis using the Wilcoxon signed rank test, with a Holm-Bonferroni correction applied to the commonly used 0.05 threshold, was used to investigate the setting relationships more closely. This analysis reveals that there is a significant difference between the errors generated by the

Table 3. A summary of the p-values from the Kruskal-Wallis test comparing the distributions of the different settings' error data for each group in both the pan and elevation dimensions.

	Pan	Elevation
lo	$p = 0.18$	$p = 0.90$
med	$p = 0.86$	$p = 0.34$
hi	$p = 0.28$	$p = 0.38$

lo and *med* settings, as well as the *lo* and *hi* settings, for Group *G1* ($p_{lo-med} = 0.003$, $p_{lo-hi} < 0.001$), showing that the *lo* setting clearly produces the highest error, while it is not clear which one of the *med* and *hi* settings is better for Group *G1*. Based on the current data, it is impossible to conclude which setting produces the smallest angular error for Group *G2*, but this may because of the relatively small sample size for each setting. Furthermore, the significant variance in the error results for Group *G2* could possibly be because of the higher mean age compared to Group *G1* and their general inexperience with mobile devices. Indeed, previous works have suggested that these demographics may have a measurable effect on target-finding performance [Millar 1994; Pring 2008]. However, comparing the distributions for each setting between the two groups with the Kruskal-Wallis test for non-parametric data, we see that the differences between the distributions for all three settings are not significantly different, for both groups and both pan and elevation (the p-values are summarised in Table 3). Further analysis did not reveal any other significant differences between the two distributions. This result show that it is not unreasonable to expect non-significantly different levels of performance should these experiments be repeated with a different group. We also note that there is a significant negative error bias in the elevation data for all the settings and groups, possibly caused by a cognitive constraint introduced by the floor, below which the participants believed the targets could not appear. Since this bias seems to be constant, it could be easily addressed in a future version of the audio interface by adjusting its frequency parameters to shift the bias upwards by a constant offset. Finally, we can conclude that the *hi* setting, which generates the significantly smallest elevation error, is the best audio pitch level to guide a user in a pointing task, and that the pan error is completely independent of such setting choice.

5.4 Time-to-Target

To investigate if the interface generates a Fitts-like response from the participants, we plot the time to find the target as a function of the targets' indices of difficulty, as defined by Eq. (2). The data is binned in intervals of the effective target width (w_e), given by Eq. (4), and are plotted for each pitch gradient setting. A logarithmic line is fitted through the bins' median values by regression and all the results are presented in Fig. 11.

For Group *G1*, a Fitts relationship can be observed and the logarithmic line of best fit closely approximates the median values of the binned data for all three settings. This is confirmed by strong Pearson correlation scores for each setting ($r_{lo} = 0.98$, $p_{lo} < 0.001$; $r_{med} = 0.94$, $p_{med} < 0.001$; $r_{hi} = 0.97$, $p_{hi} < 0.001$). Regarding Group *G2*Âÿ we observe larger spreads for each binned data interval, indicating less consistency in the time-to-target results for participants with severe sight impairments. This could be due to each participant's result being taken as a single datum and to the smaller population size in Group *G2*, as well as the mean age and general level of expertise with mobile devices of this particular group. Nevertheless, all three settings exhibit strong Pearson correlation scores ($r_{lo} = 0.71$, $p = 0.005$; $r_{med} = 0.85$, $p_{med} < 0.001$; $r_{hi} = 0.84$, $p_{hi} < 0.001$).

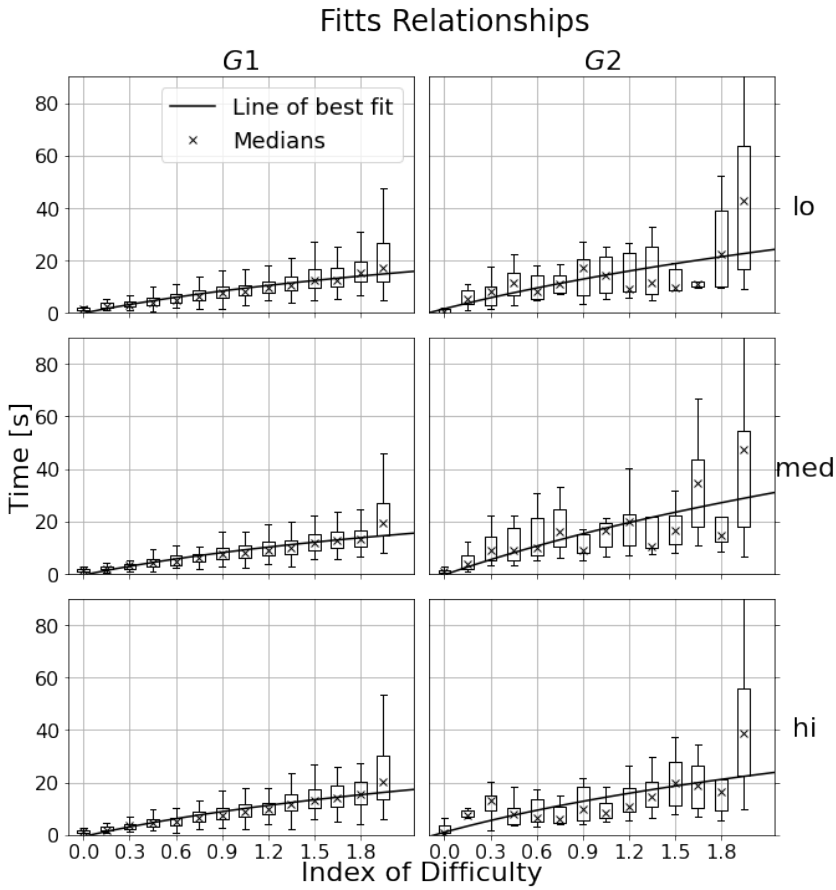

Fig. 11. Plots showing the Fitts relationship between the time it took the participants to find a target and the target's index of difficulty.

These results allows us to calculate and plot in Fig. 12 an index of performance, as given by Eq. (6), for each audio setting. The results are summarised in Table 4. For Group *G1*, there is a fairly consistent level of performance between the three settings, with *med* producing the highest indices of performance overall (i.e. the participants found the targets with the smallest error and in the least amount of time). This is supported by the results from the Friedman test, showing that there is a significant difference in performance between the settings ($p < 0.001$), as well as post-hoc Wilcoxon tests with Holm-Bonferroni corrections, which show that the *med* setting is significantly different to the *lo* and *hi* settings ($p_{med-lo} < 0.001$, $p_{med-hi} < 0.001$). The *lo* and *hi* settings, instead, are not significantly different from each other ($p_{lo-hi} = 0.11$). The results for Group *G2* show generally lower indices of performance for each setting, which is expected given the increased times to target observed in Fig. 11. Following visual inspection of the distribution and significant results from the Friedman test ($p < 0.001$), it is clear that the *hi* setting produces the highest performance by a large margin, compared to the *lo* and *med* settings. This is supported by Wilcoxon tests ($p_{lo-hi} = 0.01$, $p_{med-hi} = 0.01$), which also show that the results for the *lo* and *med* settings are not different ($p_{lo-med} = 0.09$).

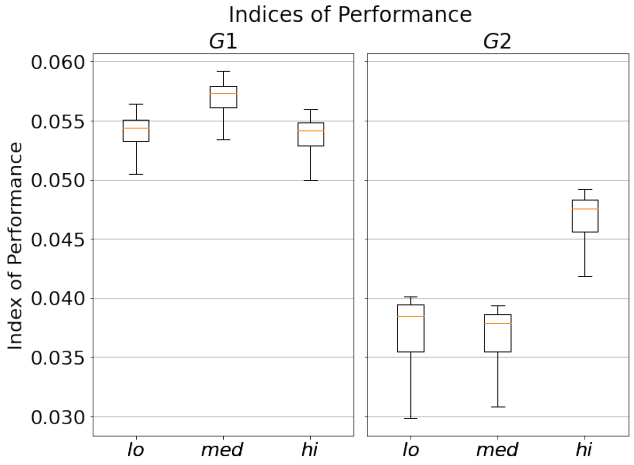

Fig. 12. Plots showing the Fitts relationship between the time it took the participants to find a target and the target's index of difficulty.

Table 4. The average target acquisition error in the pan and elevation dimensions for each participant group.

	Setting	Mean Index of Performance
G1	*lo*	0.054 ± 0.005
	med	0.057 ± 0.006
	hi	0.054 ± 0.007
G2	*lo*	0.039 ± 0.007
	med	0.038 ± 0.007
	hi	0.048 ± 0.008

Fig. 12 shows a significant difference between the IPs for each group's respective settings, with *G2* producing significantly lower indices of performance. This is further supported by a Kruskal-Wallis tests that reveals that each setting's distribution is indeed significantly different from its counterpart in the other group. The significant difference between the blindfolded group and the group of participants with visual impairments seems to indicate that the latter require significantly more time to find the target. However, it is unclear whether this is a systematic cause or a difference in search strategy between the two groups, e.g. *G2* preferring, on average, a slower and more methodical approach.

5.5 Discussion

Since the Fitts's model we used here is based on the target estimation errors, it is reasonable to expect that the accuracy and time performances will follow similar trends. Indeed, this seems to be the case, with the *hi* pitch setting producing the lowest target acquisition error, followed by the *med* and *lo* settings, respectively. This trend continues in the time-to-target results obtained with the Fitts model analysis, where the *med* setting gave the marginally highest level of performance in Group *G1* and *hi* for Group *G2*. However, the improvement of the latter settings are far clearer in Group *G2* than for *G1*. Indeed, Group *G2* seems to respond better to the increased movement

resolution that the *hi* affords the user, allowing for finer adjustments to be made to the device's orientation to get closer to the target. With the Fitts model discussed here, we can empirically evaluate any changes we make to the audio interface and optimise the parameters it to produce the desired output.

Regarding target acquisition, the progressive improvement from the *lo*, *med* and *hi* settings (see Table 2) seems to indicate that simply increasing the pitch gradient leads to better target-pointing performance. However, Fig. 7 shows that the frequency difference between the "on-target" tone and the selected one with the *hi* setting approaches the cut-off frequency of Group *G1*, indicating an inflection point where increasing the gradient reduces the final performance. Indeed, the participants from Group *G2* seem to go beyond this threshold and reach a saturation point where they can no longer reliably distinguish different tones.

6 CONCLUSION

In this paper we investigated the use of spatialised audio interface with varying pitch to guide a user with visual impairments in a target pointing task. We found that the blindfolded participants and those with severe sight impairments did not display significant performance differences in localising sound sources and differentiating between tones. Furthermore, we found that both groups did not display any significant differences in how accurately they were able to find a randomly distributed set of virtual targets. However, the blindfolded group outperformed the one with severe sight impairments in terms of time-to-target. We further tested different pitch settings and found that the user performance in the pan dimension, based on spatialised cues, is independent of such settings. Moreover, we noticed a speed/accuracy trade-off between the settings, where a higher pitch setting produces a smaller angular error, but at the cost of reducing the time performance (i.e. more time to reach the target). These results, together with an analysis done with Fitts's Law that confirms its applicability to this type of audio interface, provide a useful baseline to improve and refine the latter in future applications, prioritising speed or accuracy to produce the desired output.

This work identified a number of uncertainties that can be the focus of future work. These include questions such as what caused the observed difference in time performance between the groups and whether the constant negative bias observed in Fig. 9 is indeed caused by a cognitive bias, or whether there is a more complex underlying reason for the behaviour. Furthermore, casual post-experiment conversations with the participants revealed that some felt that one setting was easier to understand than the others and it can therefore be beneficial to investigate the possibility of adding an auto-adaptation component to the audio signal. For example, the pitch gradient can be automatically adjusted over time by the device to provide a better match between the human and computer and increase overall target-finding performance. Indeed, the work by Gallina et al. [2015] may serve as a good guideline for such a system. Additional feedback modes may also be added to allow for a clearer guidance experience for the user, e.g. vibration signals or another tone [Marentakis and Brewster 2006], to explicitly inform them when they are pointing to the target or by adjusting the volume to expand the system to three dimensions. With this kind of audio interface now better understood, it is ready to be implemented into a fully implemented guidance system for people with visual impairments.

ACKNOWLEDGMENTS

This work forms part of the ActiVis project which is supported by the Faculty Research Award: Winter 2015 from Google, while additional financial support for this work was provided by the EPSRC Network on Visual Image Interpretation in Humans and Machines (ViiHM). Finally, we would like to sincerely thank the people from the South Lincolnshire Blind Society and Lincolnshire Sensory Institute for their help in recruiting and supporting the experiments' participants.

REFERENCES

Teemu Tuomas Ahmaniemi and Vuokko Tuulikki Lantz. 2009. Augmented Reality Target Finding Based on Tactile Cues. *Proceedings of the 2009 International Conference on Multimodal Interfaces* (2009), 335–342. https://doi.org/10.1145/1647314.1647383

Aries Arditi and YingLi Tian. 2013. User Interface Preferences in the Design if a Camera-Based Navigation and Wayfinding Aid. *Journal of Visual Impairment & Blindness* 107, 2 (2013), 118–129.

Ruzena Bajcsy, Yiannis Aloimonos, and John K. Tsotsos. 2018. Revisiting active perception. *Autonomous Robots* 42, 2 (2018), 177–196.

Jens Blauert. 1969. Sound localization in the median plane. *Acta Acustica united with Acustica* 22, 4 (1969), 205–213.

J Blauert. 1997. *Spatial hearing: the psychophysics of human sound localization.* MIT press.

Jeffrey R. Blum, Mathieu Bouchard, and Jeremy R. Cooperstock. 2013. Spatialized audio environmental awareness for blind users with a smartphone. *Mobile Networks and Applications* 18, 3 (2013), 295–309.

Manuela Chessa, Nicoletta Noceti, Francesca Odone, Fabio Solari, Joan Sosa-García, and Luca Zini. 2016. An Integrated Artificial Vision Framework for Assisting Visually Impaired Users. *Computer Vision and Image Understanding* 149 (2016), 209–228.

Kai Crispien and Helen Petrie. 1994. The'GUIB'spatial auditory display-generation of an audio-based interface for blind computer users. Georgia Institute of Technology.

Barthélémy Durette, Nicolas Louveton, David Alleysson, and Jeanny Hérault. 2008. Visuo-auditory sensory substitution for mobility assistance: testing TheVIBE. In *Workshop on Computer Vision Applications for the Visually Impaired*.

Alexander Fiannaca, Ilias Apostolopoulous, and Eelke Folmer. 2014. Headlock: a wearable navigation aid that helps blind cane users traverse large open spaces. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility*. ACM, 19–26.

PM Fitts. 1954. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of experimental psychology* 47, 6 (1954), 381.

Paolo Gallina, Nicola Bellotto, and Massimiliano Di Luca. 2015. Progressive co-adaptation in human-machine interaction. *Int. Conf. on Informatics in Control, Automation and Robotics* 2 (2015), 362–368.

William G Gardner and Keith D Martin. 1995. HRTF measurements of a KEMAR. *The Journal of the Acoustical Society of America* 97, 6 (1995), 3907–3908.

Michele Geronazzo, Alberto Bedin, Luca Brayda, Claudio Campus, and Federico Avanzini. 2016. Interactive spatial sonification for non-visual exploration of virtual maps. *International Journal of Human Computer Studies* 85 (2016), 4–15.

Reginald G Golledge, James R Marston, Jack M Loomis, and Roberta L Klatzky. 2004. Stated Preferences for Components of a Personal Guidance System for Nonvisual Navigation. *Journal of Visual Impairment & Blindness,* 98, 3 (2004), 135–147.

Garin Hiebert. 2005. *OpenAL 1.1 Specification and Reference.*

Paul Kabbash and William AS Buxton. 1995. The âĂIJprinceâĂİ technique: Fitts' law and selection using area cursors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 273–279.

S. Kammoun, G. Parseihian, O. Gutierrez, A. Brilhault, A. Serpa, M. Raynal, B. Oriola, M. J.M. MacÉ, M. Auvray, M. Denis, S. J. Thorpe, P. Truillet, B. F.G. Katz, and C. Jouffrais. 2012. Navigation and space perception assistance for the visually impaired: The NAVIG project. *Irbm* 33, 2 (2012), 182–189. https://doi.org/10.1016/j.irbm.2012.01.009

Nadia Kanwal, Erkan Bostanci, Keith Currie, and Adrian F Clark. 2015. A navigation system for the visually impaired: a fusion of vision and depth sensor. *Applied bionics and biomechanics* (2015).

Brian F G Katz, Philippe Truillet, Simon J Thorpe, Christophe Jouffrais, and Jouffrais. 2010. NAVIG: Navigation Assisted by Artificial Vision and GNSS. *Workshop on Multimodal Location Based Techniques for Extreme Navigation* 1 (2010), 1–4.

Roberta L Klatzky, James R Marston, Nicholas A Giudice, Reginald G Golledge, and Jack M Loomis. 2006. Cognitive load of navigating without vision when guided by virtual sound versus spatial language. *Journal of experimental psychology: Applied* 12, 4 (2006), 223–232.

YH Lee and G Medioni. 2015. RGB-D Camera Based Wearable Navigation System for the Visually Impaired. *Computer Vision and Image Understanding* 149 (2015), 3–20.

HCCH Levitt. 1971. Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical society of America* 49, 2B (1971), 467–477.

Richard Lichenstein, Daniel Clarence Smith, Jordan Lynne Ambrose, and Laurel Anne Moody. 2012. Headphone use and pedestrian injury and death in the United States: 2004–2011. *Injury prevention* 18, 5 (2012), 287 – 290.

J.C. Lock, G. Cielniak, and N. Bellotto. 2017. Portable Navigations System with Adaptive Multimodal Interface for the Blind. In *AAAI Spring Symposium – Designing the User Experience of Machine Learning Systems*.

J.C. Lock, G. Cielniak, and N. Bellotto. 2019a. Active Object Search with a Mobile Device for People with Visual Impairments. In *Int. Conf. on Computer Vision Theory and Applications*. 476–485.

J.C. Lock, I.D. Gilchrist, G. Cielniak, and N. Bellotto. 2019b. Bone-Conduction Audio Interface to Guide People with Visual Impairments. *Communications in Computer and Information Science* (2019).

Justin A MacDonald, Paula P Henry, and Tomasz R Letowski. 2006. Spatial audio through a bone conduction interface. *International journal of audiology* 45, 10 (2006), 595–599.

IS MacKenzie. 1992. Fitts' law as a research and design tool in human-computer interaction. *Human-computer interaction* 7, 1 (1992), 91–139.

Georgios N. Marentakis and Stephen a. Brewster. 2006. Effects of feedback, mobility and index of difficulty on deictic spatial audio target acquisition in the horizontal plane. *Proc. CHI '06, ACM Press* (2006), 359. https://doi.org/10.1145/1124772.1124826

Sergio Mascetti, Lorenzo Picinali, Andrea Gerino, Dragan Ahmetovic, and Cristian Bernareggi. 2016. Sonification of guidance data during road crossing for people with visual impairments or blindness. *International Journal of Human-Computer Studies* (2016).

Susanna Millar. 1994. *Understanding and representing space: Theory and evidence from studies with blind and sighted children.* Clarendon Press/Oxford University Press.

Bogdan Mocanu, Ruxandra Tapu, and Titus Zaharia. 2016. When ultrasonic sensors and computer vision join forces for efficient obstacle detection and recognition. *Sensors* 16, 11 (2016).

Michal Pec, Michal Bujacz, Pawel Strumillo, and Andrzej Materka. 2008. Individual HRTF measurements for accurate obstacle sonification in an electronic travel aid for the blind. In *2008 International Conference on Signals and Electronic Systems*. IEEE, 235–238.

Helen Petrie, Valerie Johnson, Thomas Strothotte, Andreas Raab, Rainer Michel, Lars Reichert, and Axel Schalt. 1997. MoBIC: An aid to increase the independent mobility of blind travellers. *British Journal of Visual Impairment* 15, 2 (1997), 63–66.

CC Pratt. 1930. The spatial character of high and low tones. *Journal of Experimental Psychology* 13, 3 (1930), 278.

Giorgio Presti, Dragan Ahmetovic, Mattia Ducci, Cristian Bernareggi, Luca Ludovico, Adriano Baratè, Federico Avanzini, and Sergio Mascetti. 2019. WatchOut: Obstacle Sonification for People with Visual Impairment or Blindness. In *ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*. ACM.

Linda Pring. 2008. Psychological characteristics of children with visual impairments: learning, memory and imagery. *British Journal of Visual Impairment* (2008).

Jose Rivera-Rubio, Kai Arulkumaran, Hemang Rishi, Ioannis Alexiou, and Anil A Bharath. 2015. An assistive haptic interface for appearance-based indoor navigation. *Computer Vision and Image Understanding* 149, Assistive Computer Vision and Robotics (2015), 126–145.

Alberto Rodríguez, Luis M Bergasa, Pablo F Alcantarilla, Javier Yebes, and Andrés Cela. 2012. Obstacle Avoidance System for Assisting Visually Impaired People. *Intelligent Vehicles Symposium Workshops* (2012), 1–6.

Daisuke Sato, Uran Oh, João Guerreiro, Dragan Ahmetovic, Kakuya Naito, Hironobu Takagi, Kris M. Kitani, and Chieko Asakawa. 2019. NavCog3: Large-Scale Blind Indoor Navigation Assistant with Semantic Features in the Wild. *Transactions on Accessible Computing* (2019).

Boris Schauerte, Manel Martinez, Angela Constantinescu, and Rainer Stiefelhagen. 2012. An assistive vision system for the blind that helps find lost things. In *International Conference on Computers for Handicapped Persons*. Springer, 566–572.

David Schonstein, Laurent Ferré, and Brian F. Katz. 2008. Comparison of headphones and equalization for virtual auditory source localization. *The Journal of the Acoustical Society of America* 5 (2008), 3724–3724.

Tobias Schwarze, Martin Lauer, Manuel Schwaab, Michailas Romanovas, Sandra Bohm, and Thomas Jurgensohn. 2015. An intuitive mobility aid for visually impaired people based on stereo vision. In *Int. Conf. on Computer Vision Workshops*. 17–25.

RN Shepard. 1964. Circularity in judgments of relative pitch. *The Journal of the Acoustical Society of America* 36, 12 (1964), 2346–2353.

Raymond M Stanley and Bruce N Walker. 2006. Lateralization of sounds using bone-conduction headsets. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 50. SAGE Publications Sage CA: Los Angeles, CA, 1571–1575.

Yingli Tian, Xiaodong Yang, Chucai Yi, and Aries Arditi. 2013. Toward a Computer Vision-Based Wayfinding Aid for Blind Persons to Access Unfamiliar Indoor Environments. *Machine Vision and Applications* 24, 3 (2013), 521–535. https://doi.org/10.1007/s00138-012-0431-7

Marynel Vázquez and Aaron Steinfeld. 2012. Helping visually impaired users properly aim a camera. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*. ACM, 95–102.

Tray Minh Voong and Michael Oehler. 2019. Auditory spatial perception using bone conduction headphones along with fitted head related transfer functions. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 1211–1212.

AT Welford. 1968. *Fundamentals of skill.* Methuen.

Gyorgy Wersenyi. 2003. Localization in a HRTF-based minimum audible angle listening test on a 2D sound screen for GUIB applications. In *Audio Engineering Society Convention 115*. Audio Engineering Society.

GB Wetherill and H Levitt. 1965. Sequential estimation of points on a psychometric function. *Brit. J. Math. Statist. Psych.* 18, 1 (1965), 1–10.

Jeff Wilson, Bruce N. Walker, Jeffrey Lindsay, Craig Cambias, and Frank Dellaert. 2007. SWAN: System for wearable audio navigation. *Int. Symposium on Wearable Computers* (2007), 91–98.

Jacob O Wobbrock, Kristen Shinohara, and Alex Jansen. 2011. The effects of task dimensionality, endpoint deviation, throughput calculation, and experiment design on pointing measures and models. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1639–1648.

Jinglong Wu, Jiajia Yang, and Taichi Honda. 2010. FittsâĂŹ law holds for pointing movements under conditions of restricted visual feedback. *Human movement science* 29, 6 (2010), 882–892. https://doi.org/10.1016/j.humov.2010.03.009

Jizhong Xiao, Samleo L. Joseph, Xiaochen Zhang, Bing Li, Xiaohai Li, and Jianwei Zhang. 2015. An Assistive Navigation Framework for the Visually Impaired. *IEEE Trans. on Human-Machine Systems* 45, 5 (2015), 635–640.

Salifu Yusif, Jeffrey Soar, and Abdul Hafeez-Baig. 2016. Older people, assistive technologies, and the barriers to adoption: A systematic review. *International Journal of Medical Informatics* 94 (2016), 112–116.

Shumin Zhai, Jing Kong, and Xiangshi Ren. 2004. Speed–accuracy tradeoff in FittsâĂŹ law tasksâĂŤon the equivalency of actual and nominal pointing precision. *International journal of human-computer studies* 61, 6 (2004), 823–856.